DOCUMENT RESUME

ED 479 555 IR 022 161

AUTHOR	Jansen, Bernard J.
TITLE	Using Advanced Search Operators on Web Search Engines.
PUB DATE	2002-10-00
NOTE	9p.; In: E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education. Proceedings (7th, Montreal, Quebec, Canada, October 15-19, 2002); see IR 022 093.
AVAILABLE FROM	Association for the Advancement of Computing in Education (AACE), P.O. Box 3728, Norfolk, VA 23514. Tel: 757-623-7588; e-mail: info@aace.org; Web site: http://www.aace.org/DL/.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Comparative Analysis; Information Retrieval; Information Services; *Online Searching; Search Engines; Search Strategies; User Needs (Information); World Wide Web
IDENTIFIERS	*Query Formulations; *Query Processing

ABSTRACT

Studies show that the majority of Web searchers enter extremely simple queries, so a reasonable system design approach would be to build search engines to compensate for this user characteristic. One hundred representative queries were selected from the transaction log of a major Web search service. These 100 queries were then modified using the complex search operators supported by five major search engines. Overall, 1900 queries were submitted to five Web search engines, returning 18,256 documents. The results obtained using the complex queries from each search engine were compared to the results obtained by the original 100 simple queries from that search engine. In general, increasing the complexity of the queries had moderate effect on the results, with an average 66% exact match between results from the simple and complex queries. Implications on the effectiveness of current Web search services, for future search engine design and for future research are discussed. (Contains 23 references.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Using Advanced Search Operators on Web Search Engines

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

G.H. Marks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Bernard J. Jansen
School of Information Sciences and Technology
The Pennsylvania State University
001 Thomas Bldg
University Park, PA, 16801, USA
Phone: 814-865-6459 Fax: 814-865-6426
Email: jjansen@acm.org

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract: Studies show that the majority of Web searchers enter extremely simple queries, so a reasonable system design approach would be to build search engines to compensate for this user characteristic. One hundred representative queries were selected from the transaction log of a major Web search service. These one hundred queries were then modified using the complex search operators supported by 5 major search engines. Overall, 1900 queries were submitted to five Web search engines, returning 18,256 documents. The results obtained using the complex queries from each search engine were compared to the results obtained by the original 100 simple queries from that search engine. In general, increasing the complexity of the queries had moderate effect on the results, with an average 66% exact match between results from the simple and complex queries. Implications on the effectiveness of current Web search services, for future search engine design and for future research are discussed.

ED 479 555

Introduction

Searchers rarely use advanced techniques, such as Boolean operators or phrase searching, (Borgman, 1996) when using information retrieval (IR) systems. This characteristic has been especially true for Web searchers, with the vast majority of Web queries containing no advanced query techniques. Several Web studies have noted this near absence of complex query operators (e.g., AND, OR, NOT, must appear operators, paraphrases, etc.) in Web queries (Hoelscher, 1998; Jansen et al., 2000; Keily, 1997; Silverstein et al., 1999). The use of Boolean operators, typically about 8%, in these Web searching studies is lower than the rates reported in studies of searchers using traditional IR systems such as DIALOG or LEXIS/NEXIS, sometimes substantially. For example, Siegfried et al. (1993) reported Boolean usage of over 36% on the DIALOG system, which the researchers considered a low rate of usage.

Many IR professionals would probably say that the correct use of advanced searching operators would increase the effectiveness of Web searches. Advanced searching techniques are well known, and one can find numerous articles and books on advanced searching strategies (Korfhage, 1997), tutorials on searching training (Sullivan, 2000), and numerous educational courses on searching strategies. Many of these techniques (e.g., phrase searching and must appear operators) are easy to employ. However, based on the Web searching studies, it appears that the majority of Web searchers continue to use very simple queries, with little to no use of advanced searching operators. These searchers seem to be employing an ineffective and inefficient strategy for finding information.

This line of reasoning leads to the to the question: Why aren't Web searchers using more advanced queries? Some researchers have stated that Web searchers are just lazy (Kapur and Zhang, 2000) or that there is a design flaw with Web search engines (Clark, 2001).

Neither of these assertions seems likely. Studies and data suggest that Web users may be finding the information they want using these simple queries. A survey of users on a major Web search engine reports that

IR022161

BEST COPY AVAILABLE

almost 70% of the searchers stated that they had located relevant information on the search engine (Spink et al., 1999). Additionally, Web search engines continue to attract large numbers of Web searchers. Of the top ten Web sites in terms of number of visitors, in June 2000, eight were Web search engines (CyberAtlas, 2000), implying at least that search engines are the best alternative available for finding information on the Web. Obviously, something is amidst. Web searchers are not utilizing advanced searching operators, but they appear to be finding information using a technique that should be ineffective or at least inefficient. Although Web search engines are providing seemingly naive methods for searchers to employ, they still continue to attract a large number of customers.

To shed some light on these apparent paradoxes, we investigate the effect of complex queries (i.e., those using advanced syntax, such as Boolean operators) on the results retrieved by Web search services relative to the results retrieved by simple queries (i.e., those with no advanced syntax). Our research aim is to investigate (1) what would the effect on search engine results be if Web searchers used more advanced searching techniques, (2) whether or not the use of complex queries would be helpful for the typical Web and (3) by implication, whether or not Web search engines are somehow compensating for the simple searching techniques of Web users. We present an overview of Web searching, the study methodology, and research results from various perspectives, followed by conclusions.

Research Design and Methodology

Overview

One hundred queries were randomly selected from a transaction log of a major Web search engine. These hundred original queries, referred to as simple queries, contained no advanced query operators. Each simple query was initially submitted to a Web search engine; the query was then modified using an advanced query operator supported by that particular search engine. The results from the complex query were then compared and contrasted to the results from the simple query. The query was modified for each advanced query operator supported by the search engine. The modified queries are referred to as complex queries. The process was repeated for each of the 100 simple queries. This entire process was performed on five Web search engines.

Selection of Queries

Key to the success of this research was the selection of queries that accurately reflected the structure of Web queries. Research shows that Web queries generally have two terms (Jansen et al., 2000; Silverstein et al., 1999), cover a variety of topics (Wolfram, 1999), and are primarily noun phrases (Jansen et al., 2000; Kirsch, 1999). These criteria were used in selecting queries from an Excite search service transaction log. Specific Web query lengths reported by Jansen et al. (2000) are listed in Table 1.

Table 1: Query Length.

Terms in query	Number of queries	Percent of all queries
More than 6	1018	2
6	617	1
5	2,158	4
4	3,789	7
3	9,242	18
2	16,191	32
1	15,854	31
0	2,584	5
Total	51,453	100

The mean for query length (i.e., the number of terms in the query) was 2.21 terms with a standard deviation of 1.05 terms. These statistics are in line with those reported by other Web studies (Kirsch, 1999; Silverstein et al., 1999).

Based on the general distribution from Table 1, the queries of the following lengths were selected for this study: 10 queries of 4 terms, 31 queries of 3 terms, and 59 queries of two terms. The terms from the original queries were all lower case. Queries were selected to ensure a variety of topics since it has been reported that some search engines specialize in certain subjects (Neilsen/NetRating, 2000). Queries that appeared on the popular query lists (Searchwords, 2000) or that referenced popular entertainers, popular locations, popular songs, etc. were eliminated since Web search engines sometimes cache results from these highly queried topics (Lesk et al., 1997). For similar reasons, all queries that were obviously queries for pornography were eliminated. These goals and constraints were the guidelines for selecting the 100 queries.

Along with selecting the queries, there was also the question of what results to compare. Studies show that most Web searchers, usually about 80%, never view more than ten results (Hoelscher, 1998; Jansen et al., 2000; Silverstein et al., 1999). Based on this overwhelming evidence of Web searcher behavior, only the first ten results in the results list were selected for comparison. Moreover, we examined these results only for changes in the first ten results; relevance judgments were not made concerning the results. The ability of Web search engines to successfully retrieve relevant documents has been investigated several times (Leighton and Srivastava, 1999; Zumalt and Pasicznyuk, 1998).

Searching Environment

Many Web sites offer searching capabilities; however, this research focuses specifically on Web search engines. Search engines are the major portals for users of the Web, with 71% of Web users accessing search engines to locate other Web sites (CommerceNet/Nielsen Media, 1997). One in every 28 (3.5%) pages viewed on the Web is a search results page (Alexa Insider, 2000). Search engines are, without a doubt, the major IR systems of the Web. There are approximately 3,200 search engines on the Web (Searchengine Watch, 2000); however, only a handful of these dominate the market in terms of Web traffic. Among the better known, and those utilized in this research, are Alta Vista, Excite, FAST Search, GoTo, and Northern Light.

Searching Rules

These five search engines offer a variety of advanced searching options. Some searching options are available from each search engine's main search page. For other searching options, one must go to a 'power' searching page. For this research, only those advanced searching options available from the search engine's main page were utilized. Of the five search engines, two offer four advanced search options (+, ", AND, and OR) from the main page. Three search engines offer two advanced searching options (+, and "). All of the search engines offer dropdown boxes (e.g., language of results, document collections to search) for refining the search. When dropdown boxes were present on the main search page, the default options were utilized.

Research Structure

Each of the 100 original queries was submitted to the five search engines for a total of 500 queries. The query was then modified with the advanced searching operators supported by the various search engines. The entire process, of submitting the simple and advance queries, took 5 minutes or less. For example, the simple query *palomino horse breeder association* could be modified using the must appear operator (*+palomino +horse +breeder +association*), phrase searching operator (*"palomino horse breeder association"*), the AND operator (*palomino AND horse AND breeder AND association*), and the OR operator (*palomino OR horse OR breeder OR association*). These modified queries are the complex queries.

Results

Of the simple queries, 498 returned at least 10 results. One query returned 3 results, and one query returned no results. Therefore, there were 4983 results to use as the baseline (i.e., 498 x 10 + 3). As mentioned earlier, results that appeared higher than 10 in the results list were not utilized.

There were a total of 1400 complex queries submitted. With each search engine offering different search options, the number of complex queries varied for each search engine, as outlined in Table 2.

Table 2: Number of Queries by Search Engine and Operator.

Search Engine	Total Number of Queries	Number of Simple Queries	Number of Complex Queries	Search Options Supported and Number of Queries			
				+	"	AND	OR
Alta Vista	300	100	200	100	100		
Excite	500	100	400	100	100	100	100
FAST Search	300	100	200	100	100		
GoTo	300	100	200	100	100		
Northern Light	500	100	400	100	100	100	100
Total Queries	1900	500	1400	500	500	200	200

From Table 2, all search engines supported the must appear (+) and phrase searching (") operators for a total of 1000 queries. Excite and Northern Light support the Boolean operators AND and OR for a total of 400 queries. Submission of each set of queries on a given search engine took about 5 to 10 minutes. Therefore, the opportunity for the index on a given search engine to change during the test was limited.
Table 3 presents data on the number of results returned.

Table 3: Results from Complex Queries Results

Number of Results	Number of Occurrences	Number of Results
10	1,325	13,250
9	0	0
8	2	16
7	1	7
6	5	30
5	1	5
4	4	16
3	2	6
2	3	6
1	13	13
0	44	0
Total	1,400	13,349

Of the 1400 complex queries, 1325 returned ten or more results. There were 31 queries that returned fewer than ten results but more than zero results. There were 44 queries that returned no results. Altogether, there were 13,349 results returned by the complex queries. Combined with the 4,983 results from the simple queries, a total of 18,332 results were used in the analysis.

The match had to be exact when comparing the results between the simple and complex queries; that is, the documents listed had to be the identical page at the same site. Different pages from the same site were not counted as matches. The identical pages at different sites were not counted as matches. Furthermore, if results appeared in both lists but in a different order, they were counted as matches as long as both were displayed in the first ten results.

Simple Versus Complex Query Comparison

The aggregate results of the analysis of the 18,256 results are displayed in Table 4.

Table 4: Comparison of Simple versus Complex Queries on Major Web Search Engines.

Category	Average Number of Results that Appear in Baseline	Standard Deviation	Mode	Paired sample t
Simple Queries	9.99	0.03	10	-
Complex Queries	6.55	3.77	10	40.287
				$p < 0.01$

The baseline mean for the simple queries was 9.99, and the mean for the complex queries was 6.55. This means that, on average, 6.55 of the ten results retrieved by the complex queries also appeared in the baseline results for the corresponding simple query on that search engine. The results were analyzed using a paired sample t test, as reported in Table 4. The analysis revealed a significant difference between the two groups ($t=40.287$; $p<0.01$).

Results by Search Engine

A comparison was conducted for each search engine. These results are displayed in Table 5.

Table 5: Comparison of Results by Search Engine.

Search Engine	Mean of Simple Queries	Average Number of Results that Appear in Baseline	Standard Deviation	Mode	Paired sample t
Alta Vista	10	4.5	4.2	10	18.143
Excite	10	7.8	3.8	10	12.304
FAST Search	10	7.1	3.7	10	10.759
GoTo	9.93	8.4	2.8	10	7.704
Northern Light	10	5.4	3.9	10	23.433
		-			$p < 0.01$

From examining Table 5, we can see that Excite, FAST Search, and GoTo will on average return approximately seven to eight identical results, regardless of whether the query is simple or complex. The means for Alta Vista and Northern Light are lower at approximately four and five, respectively. Alta Vista has the largest standard deviation at 4.2. The modes of all five search engines were ten. The results from each search engine were analyzed using a paired sample t test, reported in Table 5. The analysis revealed a significant difference between the groups of simple and complex query results for all search engines.

Discussion of Results

The 100 original queries used for this research structurally represent the majority of Web queries that are more than one term. These 100 queries were submitted to five search engines. These queries were modified with advanced searching operators and also were submitted to the same five search engines. Approximately 66% of the results were identical regardless of how the searcher entered the query. Referring to the data displayed in Table 4, a paired sample t-test ($t= 40.287$, $p<0.01$) shows that the results from the simple queries are significantly different from the results for complex queries. However, as with all tests of statistical significance, one must ask what different does this make in the 'real world'?

Does it make sense to learn and utilize the more complex searching operators if on average it is only going to present the searcher with about three or four results that are different from those retrieved by just entering the query terms? Are the three or four different results worth the increased probability of entering a complex query incorrectly? As the complexity of queries increases, so does the probability of error.

The findings of this research suggest that the use of complex queries is generally not worth the trouble for the typical Web searcher (i.e., who uses two terms and is interested only in the first ten results). Based on their conduct, it appears that most Web searchers do not think it is worth the trouble either. One can say that the behavior of Web searchers adheres to the principle of least effort (Zipf, 1948), which postulates that there are "useful" behaviors that are quick and easy to perform. The very existence of these quick, easy behaviors causes individuals to choose them, even when they may not necessarily be the best behavior from a functional point of view. However, these useful behaviors are good enough, and people will generally expend the least amount of effort to achieve what they want. Zipf's theory can explain the behavior of Web searchers. The results obtained from Web search engines via simple queries are good enough. In IR jargon, the precision of simple Web queries meets the information needs of most Web searchers.

The use of simple queries versus complex queries is also compelling when one compares the modes presented in Table 4. The modes for the simple and the complex queries are both 10, meaning that more than any other occurrence, the results from a simple and complex queries will be the same.

Naturally, we acknowledge that there are cases where one is interested in all possibly relevant results. In these cases, the three or so new results may be worthwhile. Even in these cases, it would make sense to enter the simple query as well as the complex queries. Nevertheless, this type of information need, where one is interested in recall as well as precision, does not appear to be the typical information need for Web searchers. Not only are Web queries short and simple and the number of results viewed small, but the session length (the number of queries submitted during a session) of the typical Web searcher is one or two queries (Jansen and Pooch, 2000), reinforcing the notion that Web searching is a precision-based service. By their characteristics, the typical Web search does not represent a recall-based information need.

In reviewing the analysis by search engine, outlined in Table 5, there was a great deal of overlap between query results for most search engines, ranging from 45% for Alta Vista to 84% for GoTo. The mode for all five search engines was ten. Studies and presentations show that the failure rates among Web searchers using advanced syntax are high (Jansen et al., 2000). Why give searchers the opportunity to make mistakes? This rule seems to be the tactic followed by FAST Search and GoTo, which both limit the searcher's options. It also appears to be the tactic pursued by Excite where at least 78% of the results are identical, regardless of the present or absence of advanced searching operators.

These results would seem to indicate that system designers have developed the Web search engines so as to minimize the need to utilize advanced searching operators. Web searchers seldom utilize advance searching techniques, so designing IR systems to account for this characteristic seems like a reasonable course of action.

Concerning the lower mean of Alta Vista (4.5), it appears that this search engine uses a different ranking algorithm when the must appear operator (+) is used. Based on the rather random results retrieved with this operator, Alta Vista may adhere to the theoretical model of no ranking feature when Boolean-like operators are used in a query.

Conclusions and Future Research

This research indicates that use of complex queries appears to have a moderate impact on the results retrieved. On average, approximately 66% of the top ten results will be the same, regardless of how the query is entered. The approximately three or four different results on average may not be worth the increased effort required to learn the advanced searching rules or the increased risk of making a mistake. One can say that the typical Web searcher is adhering to a reasonable course of action by entering simple queries when using Web search engines.

These results also indicate that Web search engine designers are doing a proper job of designing Web interfaces and ranking algorithms. Given that the typical web searcher seldom uses advanced operators, web

search engines are to be compensating for the naïve the searching characteristics of their customers. Based on the results of this research, one can conjecture that the ranking algorithms of these search engines adhere, albeit loosely, to the following rule: *Place those documents that contain all the query terms and that have all the query terms near each other at the top of the results list.* Naturally, this is an over simplification of the complexity of searching and ranking algorithms. However, a ranking rule like this, the use of or lack of use of advanced search operators would have little impact on the results at the top of the list.

There are several avenues for future investigation. One would be to examine the effect of longer queries. The research from this study focused on the dominant behavior of Web searchers, short queries (generally about two to four terms) and viewing ten or fewer documents. It would be interesting to see if these same results hold as the query length increases. The present study would indicate that the results would not hold for query length, based on the observed decline in overlapping results in four-term queries relative to two-term queries. However, the impact in terms of explaining Web searching behavior would be less, addressing only about 7% of Web searching. Also, more discrimination would need to be made in what advanced searching operators to utilize. For example, one can make a case that the use of phrase searching might decline as the query length increased. It would also be interesting to see if these same results hold as the number of results viewed increased.

Returning our research questions, it appears that the increased usage of advanced searching techniques by Web searchers has only moderate effect on the results from Web search engines. Therefore, the use of these operators by Web searchers may not be that helpful. It also appears that the designers of Web search engines understand the searching patterns of their typical customers and are designing IR systems to compensating for the simple searching techniques of Web searchers.

References

Alexa Insider. 2000.. Alexa insider's page. Available at: http://insider.alexa.com/insider?cli=10. (30 March 2000).

Borgman, C. 1996. Why are online catalogs still hard to use? *Journal of the American Society for Information Science. 47*, 7, 493-503.

Clark, P. 2001. Solving Internet Overload. The Net Economy, 2, 3. (Note: Published electronically with no page numbers at www.thenteconomy.com.)

Commerce Net/Nielsen Media. 1997. Search engines most popular method of surfing the Web. Available at: http://www.commerce.net/news/press/0416.html (30 August 2000).

Hoelscher, C. 1998. How Internet experts search for information on the Web. Proceedings of the *World Conference of the World Wide Web, Internet, and Intranet, Orlando, FL. July 1998.* Association for the Advancement of Computing in Education. (Note: Published on CD-ROM with no page numbers.)

Jansen, B., Spink, A., Bateman, J., & Saracevic, T. 1998. Real life information retrieval: a study of user queries on the Web. *SIGIR Forum, 32*, 1, 5 - 17.

Jansen, B., & Pooch, U. 2001. Web user studies: a review of and framework for future research. *Journal of the American Society for Information Science. 52*, 3, 235 - 246.

Jansen, B., Spink, A., & Saracevic, T. 2000. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management. 36*, 2, 207-227.

Jones, S., Cunningham, S. & McNab, R. 1998. Usage analysis of a digital library. Proceedings of the *Third ACM Conference on Digital libraries, Pittsburgh, PA.* June 1998. 293 - 294.

Keily, L. 1997. Improving resource discovery on the Internet: the user perspective. *21th International Online Information Meeting, London, UK*, Learned Information Europe Ltd. 205 – 212.

Kirsch, S. 1998. The future of Internet search (keynote address). Paper presented at the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia. Available at: http://www.skirsch.com/stk.html/presentations/sigir.ppt (16 August 1999).

Korfhage, R. 1997. *Information Storage and Retrieval*. New York, NY: Wiley.

Leighton, H., & Srivastave, J. 1999. First 20 Precision among World Wide Web search services (search engines), *Journal of the American Society for Information Science*, *50*, 1, 870-881.

Lesk, M., Cutting, D., Pedersen, J., Noreault, T., & Kolli, M. 1997. Panel Session on "real world" information retrieval. *SIGIR Forum*, *32*, 1, 1 - 4.

Nielsen/Netrating 2000. Top Web Properties. Available at: http://www.nielsen-netratings.com/ (11 August 1999).

Searchwords.com 2000. Top Search Words. Available at: http://www.searchwords.com (30 May 2000).

Siegfried, S., Bates, M., & Wilde, D. 1993. A profile of end-user searching behavior by humanities scholars: the Getty online searching project report no. 2. *Journal of the American Society for Information Science*, *44*, 5, 273-291.

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. 1999. Analysis of a very large Web search engine query log. *SIGIR Forum*, *33*, 1, 6 -12.

Spink, A., Bateman, J., & Jansen, B. 1998. Searching heterogeneous collections on the Web: behavior of Excite users. *Information Research*, *4*, 2, 317 – 328. Available at: http://www.shef.ac.uk/~is/publications/infres/paper53.html (30 August 2000).

Spink, A., Wolfram, D., Jansen, B., & Saracevic, T. 2001. The public and their queries. *Journal of the American Society for Information Science*. *52*, 3, 226 - 234.

Wolfram, D. 1999. Term co-occurrence in Internet search engine queries: An analysis of the Excite data set. *Canadian Journal of Information and Library Science*. *24*, 2/3, 12-33.

Zapur, K. & Zhang, J. 2000. Searching the Web Using Synonyms and Senses. WebNet *Journal: Internet Technologies, Applications & Issues*. *2*, 3, 54 – 61.

Zumalt, J. & Pasicznyuk, R. 1998. The Internet and reference services: a real-world test of Internet utility. *Reference and User Services Quarterly*, *38*, 2, 165 – 172.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").